Using NLP and manually validated data from clinical notes to produce unbiased estimates of smoking cancer risk

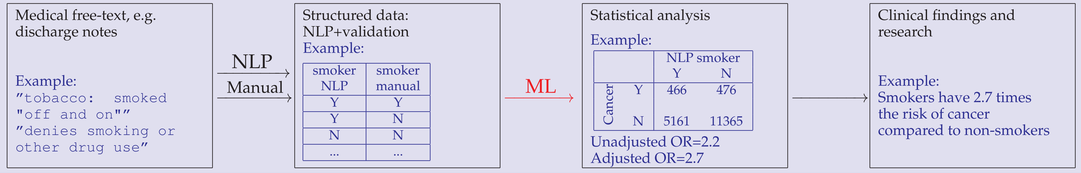
Fiona M Callaghan PhD[1], Matthew T Jackson PhD[2], Dina Demner-Fushman MD PhD[1], Swapna Abhyankar MD[1], Clement J McDonald MD[1]

[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health
[2]Office of Biostatistics, Division of Biostatistics VI, Center for Drug Evaluation and Research, US Food and Drug Administration

Motivation

- Valuable information may be contained in doctor's notes or other documents containing free-text, but it is difficult to analyze information in free-text because the information first has to be extracted and put in a structured database
- Manually extracting the information is time- and labor-intensive, and may be impractical for a large database, and Natural Language Processing (NLP) techniques provide an automated way of extracting information from large quantities of text
- However, researchers may hesitate to use information derived using NLP in a statistical analysis (e.g. to estimate risk of disease) because the information is extracted with a degree of error. When predictors are "measured with error" the resulting estimates are often biased, among other problems.[1]
- We show that our maximum likelihood (ML) method uses validation data to adjust the NLP estimates and enables the free-text information to be analyzed to produce accurate estimates of risk of disease with powerful associated inference procedures. ML method bridges the gap between the unstructured database and clinically interesting research findings.

Figure 1: Overall picture – from unstructured text to clinical results



Methods: NLP

- Rule-based smoking extraction based on limited number of features (e.g., "smok", "tobac", "cigar")[3].
- Manually reviewed a subset of discharge summaries to create dictionary of smoking-related terms.
- Positive smoking status, e.g. "smoker" and "pack-years"
- Negative smoking status, e.g. "denies smoking" and "no history of tobacco".
- Used regular expressions containing these terms to search text.

Methods: Maximum likelihood

The likelihood is the product of the likelihoods of the validation and non-validation samples. Y_i=1 if patient i has smoking-related cancer, 0 otherwise; X_i=1 if patient i is a smoker according to validation, 0 otherwise; W_i=1 if patient i is a smoker according to NLP, 0 otherwise.

$$L_v(Y,X,W) = \prod_{i=1}^{n_v}\prod_{Y_i=0}^{1}\prod_{X_i=0}^{1}\prod_{W_i=0}^{1}\Pr(Y_i|X_i)\Pr(W_i|X_i)\Pr(X_i)$$

$$L_{nv}(Y,W) = \prod_{i=1}^{n_{nv}}\prod_{Y_i=0}^{1}\prod_{W_i=0}^{1}\left[\sum_{X_i=0}^{1}\Pr(Y_i|X_i)\Pr(W_i|X_i)\Pr(X_i)\right]$$

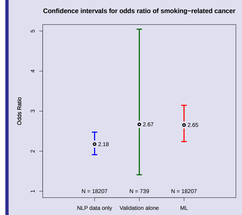
Data

Table 1: Cancer and smoking data

Smoking-related cancer	True smoker	NLP smoker $W=1$	NLP smoker $W=0$	Total
Validation sample				
$Y=1$	$X=1$	21	2	23
	$X=0$	1	17	18
$Y=0$	$X=1$	187	39	226
	$X=0$	24	448	472
Subtotal		233	506	739
Non-validation sample				
$Y=1$		466	476	942
$Y=0$		5161	11365	16526
Subtotal		5627	11841	17468
Total		5860	12347	18207

Example: Smoking and risk of cancer

Figure 2: Case Study

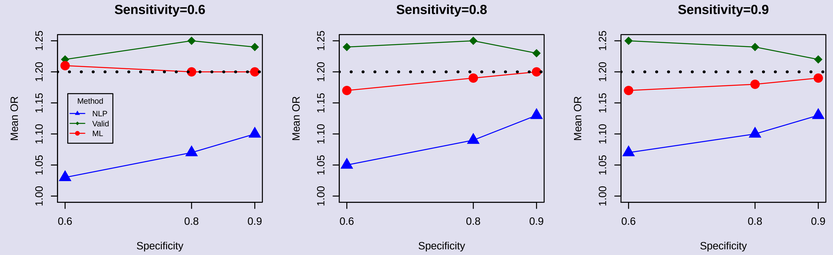


- Based on MIMIC-II [2]: an ICU database with 18,000+ patients.
- The ML OR is 2.7, ie. smokers have 2.7 times the risk of cancer of non-smokers.
- The ML estimate is close to the (true) validation estimate, but has lower variability.

Results: Simulations

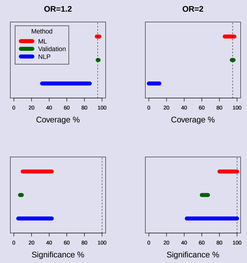
- We compared the estimates of the OR for each of the 3 methods.
- As the sensitivity increases, all methods tend to get less biased.
- The ML method (RED) performs the best, with the average simulated value closest to the true value of the OR of 1.2.
- The NLP estimates (BLUE) underestimate the true OR.
- The validation sample estimates (GREEN) tend to overestimate.
- Similar results are obtained for OR=2 (not shown).

Figure 3: Mean estimates of OR obtained from all 3 methods, as sensitivity and specificity vary, with true OR=1.2.



- The variability of the estimates is also important. Two measures used:
  - % Coverage = the percent of times the true OR is contained in the confidence interval produced by each method
  - % Significance = the percent of times the methods found a significant difference between the groups, given that there is a true difference between smokers and non-smokers when OR=1.2 or OR=2.
- For % coverage, the ML and validation sample estimates are both range around 95% – which is the correct value – with the validation sample esimates doing a little better than ML for OR=2. NLP does not have a coverage close to 95%.
- For % significance, ML method performs the best for OR=2, with the ML method finding signifcant difference between 90-100% of the time.
- For OR=1.2, none of the methods detect a significant difference more than 50% of the time.

Figure 4: Range of coverage and power of methods over all simulation scenarios



Conclusions

- The ML method can correct for the bias in NLP predictors and allow areas of free-text in a database to be analyzed using statistical prediction models, where previously the information could not be used reliably.
- Information contained in clinical notes can be extracted via NLP and used to predict risk for patients.

References

[1] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC, 2nd edition, 2006.

[2] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29:641–44, 2002. http://mimic.physionet.org/.

[3] Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24, 2008.